MECHATRONICS IN ENGINEERING DESIGN AND PRODUCT DEVELOPMENT

Dobrivoje Popovic
University of Bremen
Bremen, Germany

Ljubo Vlacic
Griffith University
Brisbane, Australia

MARCEL DEKKER, INC. NEW YORK · BASEL · HONG KONG

VISIT…

Library of Congress Cataloging-in-Publication Data

Mechatronics in engineering design and product development / [edited by]
Dobrivojie Popovic, Ljubo Vlacic.
p. cm.
Includes index.
ISBN 0-8247-0226-3 (alk. paper)
1. Mechatronics. 2. Engineering design. 3. New products. I. Popovic, Dobrivojie. II. Vlacic, Ljubo.
TJ163.12.M435 1999
621.3—dc21 98-38127
CIP

This book is printed on acid-free paper.

Headquarters
Marcel Dekker, Inc.
270 Madison Avenue, New York, NY 10016
tel: 212-696-9000; fax: 212-685-4540

Eastern Hemisphere Distribution
Marcel Dekker AG
Hutgasse 4, Postfach 812, CH-4001 Basel, Switzerland
tel: 44-61-261-8482; fax: 44-61-261-8896

World Wide Web
http://www.dekker.com

The publisher offers discounts on this book when ordered in bulk quantities. For more information, write to Special Sales/Professional Marketing at the headquarters address above.

Copyright © 1999 by Marcel Dekker, Inc. All Rights Reserved.

Neither this book nor any part may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, microfilming, and recording, or by any information storage and retrieval system, without permission in writing from the publisher.

Current printing (last digit)
10 9 8 7 6 5 4 3 2 1

PRINTED IN THE UNITED STATES OF AMERICA

Preface

In order to compete in the worldwide marketplace, manufacturers must ensure that their products will fulfill consumers' desired performance and quality requirements. Being at the synergetic intersection of a number of technologies, mechatronics has become the most instrumental tool in facilitating such company goals. Hence, many companies are gathering data on the use of mechatronic processes in engineering design and product development. Our long experience in teaching engineering at the university level and as R&D engineers with industry has led us to the belief that a practical and accessible presentation of mechatronics know-how is both possible and necessary. This book, presenting both theory and practice, is the result of this belief.

A universally accepted definition of the term mechatronics is: the integration of a number of disciplines such as mechanics, electronics, electrical, computer, control, and software engineering using microelectronics to control mechanical devices. In addition to product design, mechatronics as a design philosophy penetrates and is applied to production design, monitoring, and control with the objective of achieving high-quality products at optimal running conditions. To achieve this, mechatronics integrates advanced semiconductor technology, computer and communications technology, robotics, computer vision, and intelligent neuro-fuzzy technology.

The process of mechatronics and its interdisciplinary synergy is best explained by professionals from all the disciplines involved. This book provides systematic and comprehensive information to practicing engineers in industry and to advanced students. It also helps them to adapt this expert knowledge to their own unique situations and thus become more productive. We therefore expect that this text will serve as a reference book for professionals from the automotive, process, production and aviation industries, robotics, consumer electronics, medicine, manufacturing, and CAD centers. We also expect the text

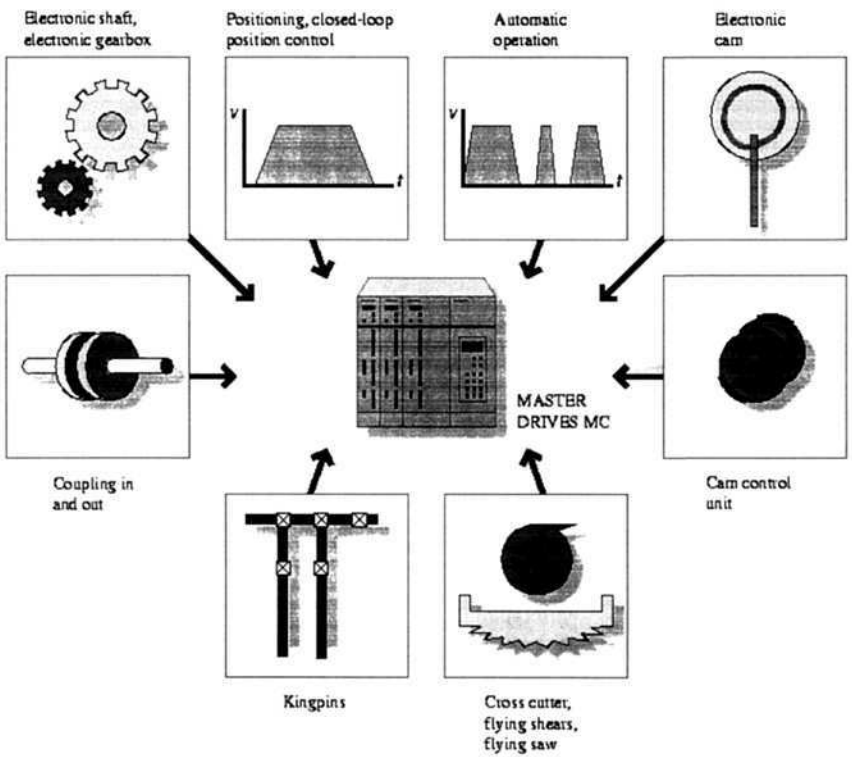
to be used in various pedagogical settings. All who are dedicated to improving engineering design and product development can make use of this book.

The design of a mechatronic product is a challenge for design engineers. It is an exciting and satisfying journey with a specific destination — that of making a product to meet the needs of the customer. While the destination is well defined, the approaches to it are different because each mechatronic product is specific in its design, with a unique composition of the disciplines involved. Having this in mind, this book aims to provide the reader with specific knowledge on how to integrate all pertinent disciplines into a mechatronic product.

Recent progress across all disciplines pertinent to mechatronics has significantly affected future mechatronic product design trends and has fast-forwarded technological enhancement of existing products. It has also led to the design of totally new classes of products such as mechatronic machines. This has been necessary because the conventional approach to machine design has limited further machine improvement due to restrictions imposed upon mechanical components and mechanical subsystems of machines. If a machine is considered to be a mechatronic system then the machine can be designed based on: (a) a minimum number of mechanical and electronic components — those hardware components that are considered absolutely necessary; (b) an intelligent unit to process all information and machine-related functions; and (c) sensors to receive and actuators to respond to this information.

The intelligent SIMOVERT Master Drives Motion Control Servo Converter of Siemens, Germany, is being designed in line with mechatronic machine design principles. As Siemens describes it: "To start with, each mechatronic machine will have its own drive. Mechanical coupling such as cams and kingpins will be eliminated. The electronics will precisely coordinate the various movements. Motion sequences will also be coordinated. Each drive will know what the other drives are doing based on interlinks. Typical applications such as start-up, positioning, synchronous operations and cam can be called upon as standard functional blocks. All these functions, now performed by software, are to be configured to suit a specific machine application. The response so far from the end users of pioneering mechatronic machines is that implementation of the mechatronic approach to machine design has reduced the operating and maintenance costs of the machines and increased their application flexibility."

Mechatronic process is therefore a cross-disciplinary design process which can be properly applied if, and only if, the specialists from all pertinent disciplines work together from very early in the design process. Marketing specialists and production engineers should also participate. This interdisciplinary team-based work is usually called concurrent engineering due to the simultaneous interrelationships of the disciplines and their influence on the final solution. A mechatronic product is therefore not only a simple collection but a synergistic composition of all pertinent disciplinary knowledge.



Frequently used applications are integrated into the drive converter as standard mechatronic functions. (Courtesy of Siemens, Germany.)

It is up to the design team to find and define the composition, harmony and balance of the disciplines involved.

With the above in mind, this book is presented in four parts. Part One of the book addresses the core technologies necessary for the design and development of the mechatronic product. Chapter 1 discusses the transducers (sensors and actuators) most commonly used in mechatronics with the main emphasis on sensors, especially position/displacement sensors. Chapter 2 is dedicated to the advanced solutions recently developed in the area of microsensors and microactuators.

Chapter 3 describes the design philosophy of the microcontroller, a form of microcomputer, which is the intelligent core of a mechatronic system and is responsible for processing information received by the mechatronic product via its sensors. This chapter focuses particularly on designing microcontroller and associated circuits for a target system and discusses prototype implementation techniques. Three prototype design case studies are explained in detail.

The scope of real-time information processing, which is required to be performed by the mechatronic product, may comprise simple measurement and/or control functions but also may consist of complex supervisory, optimization, knowledge-based, and intelligent control functions. The theory behind the machine's ability to be intelligent is explained in Chapter 4.

Part One ends with Chapter 5, an introduction to communications technology. This knowledge is considered essential when integrating the mechatronic product. Some specific bus systems, local area networks, and related inter-networking elements such as network bridges and network gateways are presented and their applications discussed.

Part Two deals with some design approaches, including conceptual design, and relies on the distributed structure of production systems, summarized in Chapter 6. Chapter 7 shows how an automotive engine controller can be developed using computer-aided design tools. A step-by-step explanation of the whole computer-aided design process shows how the design can be coded into the target processor and tested. Chapter 7 also describes the state of the art for automatic code generation, analysis, and synthesis of mechatronic systems that are controlled by computers.

If the controller for a mechatronic product is to be designed, developed, and prototyped in hardware as an Application Specific Integrated Circuit (ASIC), the know-how needed can be found in Chapter 8. The chapter presents an overview of methods, tools, and the latest technology used in the rapid prototyping of mechatronic systems. The main emphasis is on design tools for the rapid prototyping of the mechanical and electronic components of a mechatronic product. Both of these components need to be rapidly prototyped early in the design stage of mechatronic product development in order to evaluate the performance expected to be fulfilled by the target system, to evaluate likely customer response, desired reliability, and so forth.

Related mechatronic product design aspects are grouped in Chapters 9, 10, and 11, which belong to Part Three of the book. In Chapter 9, the design aspect of system integration, optimality, and compatibility of the system elements is presented. Guidelines to the selection and interface of the system elements and the measurement of resulting reliability and robustness of the integrated system are also provided. In Chapter 10, system performance aspects are discussed, with particular attention to production and product quality monitoring, quality assurance, and control. An issue presented within Chapter 11, system software, is the crucial design issue in relation to the real-time application of mechatronic products for which effective interaction between the system and its immediate environment is essential to the performance required of the product.

Part Four, consisting of three chapters, addresses some mechatronic products application-related issues. Chapter 12 describes the versatility of mechatronic system applications. Among the case studies, explained are mechatronic development in gear measuring technology, automatic calibration system for an angular encoder, a construction robot for marking of the ceiling,

and, finally, musician robots playing a trio (recorder, violin, and cello) of chamber music. Chapter 13 discusses control and optimization of mechatronic processes describing an operator's model which is applicable to many complex industrial processes that require human intervention. Application examples derived from a pH neutralization process and gear ratio control problems are discussed. These demonstrate applicability of the model to a variety of industrial, manufacturing, and other dynamic large-scale processes where the operator is called upon to exercise corrective actions based upon experience. Chapter 14 is dedicated to the ethics of product design and introduces the reader to the realm of ethical problem solving, emphasizing the similarities between it and the design process with which most engineers are familiar. It shows that ethical considerations can both drive and constrain engineering design and concludes with an examination of a case directly related to the field of mechatronics.

Our profound gratitude goes to all chapter authors. Without their enthusiasm and strong dedication, the manuscript for the book would not have been completed nor its high quality achieved.

Having done our best to review the manuscript carefully, we share the blame for any shortcomings. We give full credit to the authors for the value of their contributions that enabled us to bring this volume to the community of engineers and scientists involved in mechatronic product design and development.

When all is said and done, the book could not have been produced without the able assistance of the production editors, at Marcel Dekker, Inc., Matthew MacIsaac and Brian Black. Their patience and expert guidance were instrumental in converting the manuscript into a book. For this, we are sincerely appreciative.

Dobrivoje Popovic
Ljubo Vlacic

1

Sensors and Actuators in Mechatronics

Wanjun Wang
Louisiana State University, Baton Rouge, Louisiana

In the past several decades, partly because of the rapid development of the microelectronics industry and the ever-increasing applications of microcomputers and the automation of various industries, demands for transducers (sensors/actuators) have increased exponentially. This trend is expected to continue as global competition for higher productivity and better quality forces companies in every industry to be constantly looking for ways to reduce cost and improve the quality of their production. In one respect, the revolution in microelectronics and computers has dramatically reduced the cost of automation and control, and has therefore led to broad applications of these technologies in areas where these technologies were not deemed to be economically feasible. On the other hand, the developments in the microelectronics industry and, more recently, the fast development of microelectromechanical systems (MEMS), have generated a wide variety of sensors and actuators at ever-lowering costs, therefore opening up opportunities for applications in some areas not feasible in the past. The fast development of microelectronics has also dramatically improved signal processing and computation capabilities. A signal processing job that might have required a large box of electronic components decades ago can now be carried out by a single IC chip. The availability, the simplicity, and the performance of circuit modules such as active filters, analog dividers, sample/hold devices, function generators, lock-in amplifiers, etc., have made it very convenient to integrate sensors and actuators in a mechanical system and has made the lives of application engineers much easier. As a consequence, integration of sensors and actuators into a system and dealing with signal conditioning, work that was deemed to be the province of an electrical engineer many years ago, can now be done by mechanical, civil, or chemical engineers. This helps expand further the application of advanced sensors in industry.

Nowadays, the application of sensors is so pervasive that it is difficult to find any machines or appliances that do not have integrated sensors. A typical car now has more than 70 sensors and the number grows continuously as efforts towards better performance are made. As the trend towards intelligent vehicles continues, more advanced or smarter sensors will be implemented, for example, ultrasonic sensors for collision prevention, bar-coded intelligent highways and vehicles with optical scanning sensors and computers as well as vehicles with built-in global satellite positioning systems and other sensors for location and guidance. Other common examples of mechatronics products include home appliances such as washing machines and dryers. They are no longer the very simple appliances as they were years ago and have become intelligent machines with many sensors and functions. Go to any toy store, and you will be amazed to find how many of the toys have integrated sensors and are really intelligent toys.

Transducers (sensors and actuators) are to mechatronics systems as the sensing organs and hands and feet are to human beings. As a matter of fact, in most cases the overall performance of a system is set by the performance of the sensors and actuators used. In an instrument, the sensor transforms the physical parameter to be measured into a signal as shown in Figure 1. In most modern instruments, the physical parameter being measured is transduced into an electrical signal. This signal is then processed by a signal conditioning circuit and displayed on a panel or stored for future use or processing. Obviously, the overall sensitivity of the instrument will not be higher than that of the sensor used regardless of the quality of the remaining parts of the system.

The same argument holds for a closed-loop control system. Take a position control system as shown in Figure 2 for example. A target position is set by the operator. Based on the control strategy adopted, a command is generated and sent to the actuators which then drive the plant to the target position. The output position is then measured with a sensing system and compared with the target position. If there is a difference between the measured and the target positions, a correction command is generated and sent out to compensate for the error. This check-correction action continues until the target position is reached. The revolution in microelectronics and microcomputers has made the implementation of advanced control systems and high quality electronic components readily available at ever-lowering costs, and they have therefore become a less vital part of an automatic system. Consequently, researchers and engineers have found that the overall performance of the system is nearly always limited by the performance of the sensors and actuators, especially the sensors. If the sensing system cannot

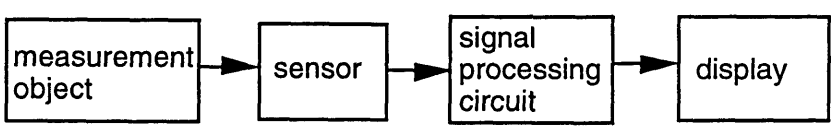


Figure 1 Schematic diagram of an A instrument.

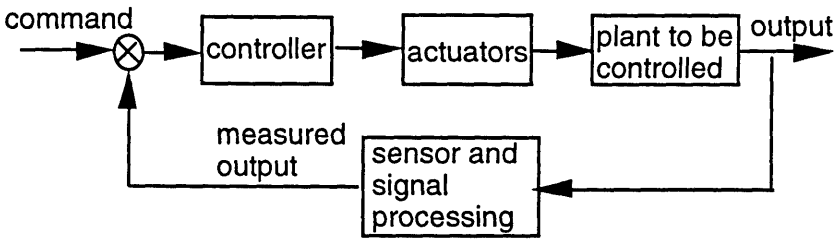


Figure 2 Schematic diagram of an A closed-loop control system.

accurately measure the actual position output, the wrong control command will be generated no matter how fancy the controller is. Similarly, if the actuators are not accurate enough, the correct control commands can still lead to an erroneous output. Also, the dynamic characteristics of the whole system are affected by the characteristics of the transducers used. Therefore it is very important to study the dynamic behavior of the sensors and actuators.

In this chapter, we are going to discuss the transducers (sensors/actuators) most commonly used in mechatronics, with the main emphasis on sensors, especially position/displacement sensors. In the following sections, classification of sensors and actuators will be introduced, followed by the basic concepts and common terms used in sensors and actuators. Then the fundamental principles of several major categories of transducers will be presented. In particular, two of the most commonly used position/displacement sensors, LVDTs and PSDs, will be discussed in detail.

I. CLASSIFICATION OF SENSORS AND ACTUATORS

A. Two Categories of Transducers: Sensors and Actuators

The term *transducer* is widely used in science and engineering. However, there are several different definitions, each having its own advantages and disadvantages, and they are all used in the field. Middlehoek [1] defines a transducer as a device that transforms non-electrical energy into electrical energy or vice versa. Karnopp et al. [2] defines a transducer as a device that transforms energy from one domain to another. The last, and the broadest definition was proposed by Busch-Vishniac [3]: a transducer is a device that transforms energy from one form to another, and it does not matter whether the energy belongs to different domains or the same domains. For example, a load cell, which is commonly used in mechanical measurement, can be classified as a transducer even though both the input energy and output energy are in the mechanical domain. This definition is more general than the one suggested by Middlehoek. If we insist a transducer must have an electrical output, a regular thermostat will not be classified as a transducer because it transforms thermal energy into mechanical energy and no electrical energy is involved. A load cell will not be classified as a transducer using this criterion. Strictly

speaking, even a strain gauge, one of the widely used mechanical sensors, cannot be called a transducer (sensor) because it only transforms the mechanical energy (strain) into a change in resistance, not an electrical signal.

Transducers are normally classified using two categories: sensors and actuators. In most cases, they are reciprocal. For example, a capacitive transducer may function as a position sensor that transforms mechanical energy into capacitance change, and therefore electrical signal output (electrical charge or voltage change). It may also function as an actuator when an electrical signal (voltage or charge) is supplied and an electrostatic force or displacement is delivered as output. Both cases have practical applications: for example, the rangefinder of the auto-focusing system for a popular type of Polaroid camera [4]. This camera uses an electrostatic type transducer as both an actuator to generate the ultrasonic pulses and a sensor to receive the acoustic signal echoed back. Similarly, a piezoelectric transducer can either be used as an actuator, transforming an input voltage signal into a controlled displacement (or to deliver a driving force) as commonly used for precision positioning applications such as driving atomic force microscope (AFM) tips. When a force (or pressure) is applied on a piezo-electric disk, a voltage signal is generated across the piezo-electric disk. By measuring the voltage signal, the force or the pressure can be measured. This principle is used for pressure measurement and in accelerometers. Of course, not every type of transducer can be used with equal effectiveness in sensing and actuation. As a matter of fact, optimization of the performances of sensors and actuators often calls for totally different, or most possibly, conflicting requirements in choosing design parameters. In general, the optimization of a sensor requires: (1) a minimum change in the object parameter (called *measurand*) can cause a maximum change in the state of the sensor output; (2) the influence of the sensor on the object being measured is minimal, that is, the status of the object being measured should not be affected by the presentation of the sensor itself; and (3) the output of the sensor is only affected by the desired input, not by any other parameters or environmental conditions. For example, when a thermostat is used to measure the temperature of an object, it is definitely not desired that the temperature of the object is significantly affected by the presence of the thermostat, or that the thermostat readout is influenced by the surrounding magnetic field. Strictly speaking, a thermostat will always change the temperature of the object being measured from the basic principle of the measurement. A thermostat absorbs heat from the measurement object when they come into contact; its temperature goes up as a result of this heat transfer. Therefore a large sized thermostat is always avoided because large physical size almost always means large thermal capacitance, and consequently a significant amount of heat needs to be transferred from the measurement object to the thermostat to bring the thermostat to the same temperature as the object being measured. This significant heat transfer means that the temperature of the object will inevitably be changed by the presentation of the thermostat. The smaller the thermal capacitance of the thermostat, the less significantly the temperature of the object will be changed. This also explains why sensors with smaller physical sizes

and weight are always preferred in engineering practice. For an actuator, optimization calls for them being able to impose the desired state on an object regardless of the load applied to them. Taking a DC motor as an example, the ideal performance for the motor means a motor should be able to drive anything connected to it in a specified speed independent of the load it has to overcome. This means an unlimited supply of power is required, a case that can never be achieved. Because actuators are always required to deliver certain power, their physical sizes will be power-limited in most practical cases. In most cases, actuators of extremely small size will not be very useful because of their very limited power capability.

From the foregoing discussions, it is obvious that smaller physical size and lower power requirement is always preferred for sensors while for actuators this is not always true. Partly because of this, progress made in the field of microelectromechanical systems (MEMS) and solid state technologies has greatly advanced the state-of-the-art for sensors while very limited success has been achieved in developing actuators.

B. Classifications of Sensors and Actuators

Because of the great variety of sensors and actuators used in the field of mechatronics, it is very difficult to discuss all of them in detail in the limited space available here. In this section, a very general classification of the most commonly used sensors and actuators in mechatronics will be provided, with emphasis on sensors. The classification of transducers can be done in several different ways. The first approach may be to classify sensors according to their applications, or the physical quantities the sensors can be used to measure. This is probably the most popular method of classification. Using this method, the sensors can be classified as, for example: position or displacement sensing; pressure sensing; temperature sensing; magnetic field sensing; flow measurement; torque sensing; stress or strain sensing; gas sensing; humidity sensing; chemical sensing; biological sensing; velocity or acceleration sensing; acoustic sensing (e.g., sound intensity sensing); radiation sensing. Some representative categories of sensors are shown in Table 1. One major advantage of this classification is that all the sensors for the same application can be introduced as one category and compared by performance and limitations, which can be very convenient for application engineers. There are several major disadvantages in using this classification. First, this method of presentation may become simply a review of what is commercially available today, which is not very useful to engineers who may want to study the fundamentals in sensors and actuators and learn how to develop similar new sensors. Secondly, sensors for same applications may be based on totally different principles, therefore in-depth discussions may lead to unnecessary repetitions and cannot be organized easily into the limited space here.

Another commonly used approach is to classify the sensors according their basic operation principles: optical sensing; capacitive sensing; inductive sensing; acoustic sensing (e.g., ultrasound sensing devices); fiber-optic sensing; Hall-effect

Table 1 Classification of Sensors According to Their Application

Mechanical sensors	Position (linear and rotational), displacement (linear and rotational), velocity (linear and rotational), acceleration (linear and rotational), vibration (linear and rotational), stress and strain, force, torque, pressure, surface topography or roughness and flatness, roundness, etc.
Electrical and magnetic sensors	Voltage, current, resistance, capacitance, inductance, magnetic, radiation, etc.
Acoustic and flow sensors	Sound intensity (pressure), viscosity, flow rate, frequency, ultrasound nondestructive detection, etc.
Chemical and biological sensors	pH, enzymes, ions, gases, concentration, humidity, biological, frequency shift or Doppler, etc.
Optical sensors	Intensity, wavelength, phase, vision and image (e.g., CCD camera), interference, polarization, reflectance, transmittance, scattering, refractive index, spectrum, etc.
Thermal sensors	Temperature, infrared radiation image, etc.

sensing; eddy current; and etc. as shown in Table 2. It should be noted that the list in Table 2 is only representative, not inclusive. There are some sensors not listed because of the limited space here. For example, resonant sensors, magnetoresistive sensors, etc. This method of classification is convenient for in-depth discussions of each category of sensors and is very suitable for efficient presentation of the wide variety of sensors within the confines of this chapter.

We will concentrate our discussions on sensors. Actuators will be discussed only when they are reciprocals of a particular type of sensor. Because of the limited space here, very brief discussion will be provided for each category of sensors. In each category of sensors, only one or two examples will be provided. In addition, because it is estimated that almost 80% of the sensors used in industry are for position measurement, see Luo [5], we will spend most of our efforts on studying position/displacement sensors. In particular, we will discuss in detail two of the best position sensors: linear variable differential transformers (LVDTs) and lateral effect position sensitive detectors (PSDs).

II. PERFORMANCE PARAMETERS OF A SENSOR

In this section, we will briefly discuss the terms commonly used in science and engineering practice to describe the performance of sensors. The databooks of commercial sensor products tend to use different terms to describe the same parameters of products and this makes it quite difficult to compare their performance. To avoid possible confusion, we will first define the fundamental terms used in this chapter. Efforts have been made to define these terms consistently with common usage.

Table 2 Classification of Sensors to Their Fundamental Principles

Capacitive	Position (linear and rotational), displacement (linear and rotational), electrostatic driving and deflection sensing for MEMS devices, chemical sensing, etc.
Inductive	Electric sensors, position/displacement, proximity, magnetic field detection, electromagnetic relay, etc.
Ultrasonic	Range-finders, nondestructive testing, thickness measurement, image scanning, flow measurement (Doppler), etc.
Photoelectric, PSD, CCD	Displacement/position, temperature, vision, and image (e.g., CCD camera), light intensity
Optical and fiber-optic	Optical encoders, gyrator (fiber-optic), temperature, magnetic, fiber-optic interferometer for phase shift measurement and any physical parameters that can modulate the phase shift, proximity sensors, etc.
Eddy current	
Hall effect	Magnetic field detection, proximity sensor
Piezo-electric	Actuators, pressure, force, and torque sensing, etc.

Generally speaking, in choosing sensors we must decide what the sensor is to do and what results we expect. We will discuss some of the criteria that must be considered in selecting and using different kinds of sensors for different kinds of applications in an automation system. The most important parameters are defined and discussed in the following subsections.

A. Sensitivity

Sensitivity is defined as the ratio of change of output to change in input. Suppose the output of a transducer is y for a given input x, that is, $y = f(x)$. This ratio is:

For example, if a 0.0l mm displacement in input gives rise to a 0.5 volt

$$S = \frac{\Delta y}{\Delta x} \tag{1}$$

For example, if an 0.0l mm displacement in input gives rise to an 0.5 volt change in output, then the sensitivity is 50 volt/mm. Some people prefer to use sensitivity to indicate the smallest input that can be detected by the sensor, but in most cases, another term, *resolution,* has been used for this purpose. Normally, the maximum sensitivity is always desired if other parameters such as *linearity* and *accuracy* would not be sacrificed.

B. Linearity

The term *linearity* is used to indicate the constancy of the ratio of output to input. If the output and input of a sensor system have a perfectly linear relationship, it would mean that in the following equation:

$$y = cx \tag{2}$$

where y is output, x is input, and c would be constant. If there exists some nonlinearity, c would be a function of x. That is, instead of the previous equation, the actual relationship between the output and input of the sensors would be described by the following nonlinear equation:

$$y = f(x) \tag{3}$$

There are several ways to describe linearity. One way to measure linearity is to use the absolute maximum error between the output predicted by using the linear equation and the actual output over the whole working space:

$$distortion = Max\left|f(x) - cx\right| \tag{4}$$

Here *distortion* is used to indicate the linearity. The smaller the distortion, the higher the linearity is. Another way to describe the linearity is to use a relative value for *distortion:*

$$distortion = \frac{Max\left|f(x) - cx\right|}{cx} \tag{5}$$

This definition is fairly popular and has been adopted in the discussion throughout this chapter.

C. Range

Range is a measure of the difference between the maximum and minimum values measured. For example, a thermometer might be able to measure values over the range of -40 degrees centigrade to 100 degrees centigrade. A large working range is always preferred. Normally there is always some kind of trade-off between the range and accuracy. It is generally quite expensive to achieve both high accuracy and large range in a sensor.

D. Accuracy

Accuracy is the term used to indicate the difference between the measured and the actual values. It can either be defined as an absolute value which is the maximum error within the working range, or a relative value which is the ratio of the maximum error to the range of measurement and is a dimensionless value. An accuracy of ±0.01 mm means that, under any circumstances, the value measured by the sensor would be within 0.01 mm of the actual value. Accuracy is a specification very hard to check. For example, if a robot end-effector is claimed to have an accuracy of ±0.01 mm, to verify this accuracy would require very careful measurement of the end-effector with respect to the base coordinates under specified operating conditions with regard to speed, temperature, force, torque, and load. The overall error must be within the specified range. Accuracy is one of the most rigorous technical specifications.

E. Repeatability

Repeatability is used to indicate the difference in value between two successive measurements under the same environmental and operational conditions. Repeatability is specified in most robot systems. Compared with accuracy, it is a far less stringent criterion. In most cases, no matter how poor the other parameters are, a relatively better repeatability can be expected even if the operational and environmental conditions are maintained.

F. Resolution

Resolution is defined as the minimum change of input that can be detected at the output of the sensing system, or the minimum change of the output parameter that can be achieved for an actuator. For a position sensor, the resolution would be the minimum displacement that can be detected. For a linear motor (a commonly used actuator), the resolution is defined as the minimum position displacement it can be controlled to move. Some people also prefer to define resolution as the number of measurements within the range from minimum to maximum. Resolution is one of the most important parameters of sensors and actuators.

G. Output

The type of *output* is also very important for sensors. It can be electrical current or voltage. It can also be a mechanical movement, a pressure change, liquid level variation, or resistance change. In most of today's applications, a voltage output signal is preferred because computers are being increasingly used to control the systems. If the output signal is not a voltage signal, it is often converted to a voltage signal.

H. Dynamic Characteristics

All the performance parameters discussed above are steady-state characteristics. However, in most mechatronics applications, in addition to the requirement for satisfactory steady-state characteristics, a transducer also must have satisfactory dynamic characteristics.

Bandwidth

Bandwidth is one of the most important dynamic parameters of a transducer. Ideally, we would prefer a transducer to have the same amplitude of output for any input signals with the same amplitude, independent of their frequencies. However, this ideal case can never be achieved. In reality, the output of a transducer is dependent on both the amplitude and frequency of an input signal because of the limited bandwidth of the transducer. Both the amplitude and the phase of the output signal can be a function of the amplitude and frequency of the input signal. This dynamic characteristic can be better discussed in the frequency domain, in term of the *transfer function.* Assuming a sinusoidal input signal is supplied to a

transducer, we should expect the output of the transducer to be a sinusoidal function. The ratio of the output and input signals is therefore defined as the *transfer function* of the transducer. Therefore the transfer function of the transducer is a function of the frequency of the input signal. If the amplitude of the transfer function is plotted out as the function of the input frequency it should, in most cases, look like the response curve shown in Figure 3. Below a certain frequency limit, this amplification factor of the transducer, $|H(\omega)|$, should be nearly independent of the input frequency α. As the frequency of the input signal increases, the amplitude of the output signal will decay. If the transducer can be modeled as a second (or higher) order system, there may exist a resonant frequency, or multiple resonant frequencies.

The *bandwidth* of a transducer is defined as the frequency range in which the amplification factor or the amplitude of the transfer function will not decay significantly. In Figure 3, the bandwidth is α_B.

Response Time

Another term frequently used to describe the dynamic characteristics of a transducer is its *response time*. In control theory, the response time is called settling time. The response time can be defined as the time required for a change in input to become observable as a stable change in output. In most cases, the output of the sensor would oscillate for a short time before it finally reaches and stays within a specified percentage of the final (steady-state) value. In these cases, we measure response time from the start of an input change to the time when the output has settled to the specified range. The response time of the sensor system is determined by its time constant.

III. CAPACITIVE SENSORS AND ACTUATORS

In this and following sections, a unified energy approach will be adopted in the derivation of all the fundamental equations. This energy approach has been

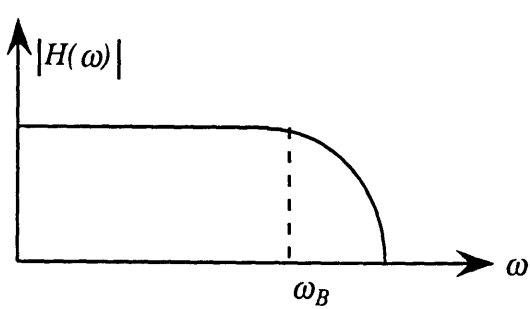


Figure 3 The bandwidth of a transducer.

used extensively in the theory of physical system dynamics by Karnopp et al. [2] and has been applied in systematic modeling and analysis of electromechanical sensors and actuators by Busch-Vishniac [3]. Because of the limited space here, the discussion will follow Busch-Vishniac's approach to this type of problems and will be brief. Interested readers are referred to the book by Busch-Vishniac [3].

The electrical charge q, the capacitance C, *and* the electrical voltage across the two plates of a capacitor e, are related by the following equation:

$$q = C \cdot e \tag{6}$$

For a parallel plate capacitor, the capacitance can be derived as the function of the distance d between the two plates, the cross-sectional area A of the plates, and the dielectric constant M as:

$$C = \frac{\varepsilon A}{d}. \tag{7}$$

Suppose one plate of the capacitor is stationary and another is moving as shown in Figure 4A and the gap between the two plates is x, the effective cross-sectional area is $A = w \cdot y$, where w is the width of the capacitor plate. The electric energy stored in the capacitor can be obtained using

$$E = \frac{1}{2C} q2 = \frac{xq2}{2\varepsilon A} = \frac{xq2}{2\varepsilon wy}. \tag{8}$$

This expression for energy can then be differentiated with respect to time to obtain an expression for the power involved as

$$P = \frac{dE}{dt} = \frac{\partial E}{\partial q}\frac{dq}{dt} + \frac{\partial E}{\partial x}\frac{dx}{dt} + \frac{\partial E}{\partial y}\frac{dy}{dt}. \tag{9}$$

Because the left side of the equation is power, and dq/dt is electrical current, then the first term $\partial E / \partial q$ must be the voltage. Similarly, the fact that dx/dt and dy/dt are velocities of the moving plate in the horizontal and vertical directions respectively means that the other two terms, $\partial E / \partial x$ and $\partial E / \partial y$ must be the mechanical forces in the x and y directions respectively in order to produce power. Therefore, we can obtain expressions for the voltage e across the capacitor and the forces F_x and F_y required to drive the moving plate of the capacitor in the x and y directions respectively, using:

$$e = \frac{\partial E}{\partial q} = \frac{qx}{\varepsilon wy}, \tag{10}$$

and

$$Fx = \frac{\partial E}{dx} = \frac{q2}{2\varepsilon wy} = \frac{\varepsilon wy}{2x2} e2, \tag{11}$$

and

$$F_y = \frac{\partial E}{\partial y} = \frac{q2}{2\varepsilon w y2} = -\frac{\varepsilon w}{2x} e2 . \tag{12}$$

Equation (10) shows that this parallel plate capacitor can be used for sensing purposes. There are several possible sensing mechanisms: (a) with one plated fixed and another moving vertically (with y = constant), the voltage e across them is then the function of the gap x between them as shown in Figure 4A; (b) with the gap between the two plates fixed (with x = constant), the bottom plate fixed and the top one moving laterally, then the effective area $A = w \cdot y$ changes as a function of the lateral position y as shown in Figure 4B; and (c) with both plates fixed (no vertical and lateral movement permitted) the effective dielectric constant of the medium can be changed by either moving a solid dielectric material between two capacitor plates as shown in Figure 4B. An example of sensors based on the principle shown in Figure 4B is a rain gauge or fluid level sensor. As the fluid (for example, water) level rises, the equivalent dielectric constant changes, and therefore the voltage across the two plates changes. This last scheme may also be used for chemical sensing because different materials may have different dielectric constants.

Equations (11) and (12) show that the parallel plate capacitor may also be used as an actuator. When a charge is supplied (and a voltage generated), either a vertical force F_x (perpendicular to the two plates) or lateral force F_y can be generated. Of course, the scheme shown in Figure 4B can also be used. In this scheme, a solid block of dielectric material is sliding in and out between the two capacitor plates; the capacitor can be treated as two capacitors with different dielectric materials in parallel. Its total energy needs to be differentiated with respect to y to obtain the force.

The capacitive sensing and driving principles have been widely used in the field of microelectromechanical systems (MEMS). Various types of microdevices and systems have been developed based on this principle. A typical example is the microfabricated pressure sensor that uses a capacitive sensor, see ref. [6]. Another example is a microvalve with electrostatic actuation described by Bosch et al. [7]. The foregoing discussion is only illustrative, and not exclusive. Other types of capacitive sensors are used also widely in engineering practice. Obviously a cylindrical capacitor with a hollow tube as one plate and a solid rod or hollow tube

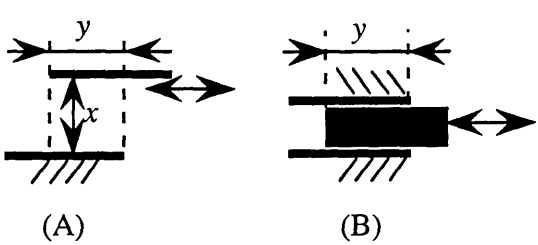


Figure 4 Several ways to use a parallel plate capacitor as a sensor.

with smaller diameter sliding inside as the other plate may also be used. The basic function is still $q = C \cdot e$ as in the case of the parallel plate capacitor, only the expression for the capacitance C is changed. The basic equations governing it can also be derived with the same energy approach as the one just shown for the parallel plate capacitor.

Another example of capacitive position sensors is a proximity sensor. In a proximity sensor, the two capacitor plates are in the same plane instead of in parallel. As the distance between the two plates and the object changes, the effective dielectric constant between the two capacitor plates changes; then the capacitance measured will be modulated by the position of the object. Proximity sensors based on this principle are commercially available.

The major advantage of capacitive position sensors is high precision. The main disadvantage of them is their small working range. They are widely used in vibration and displacement measurement. They are also one of the most commonly used transduction principles used in the fast developing MEMS field.

IV. INDUCTIVE SENSORS/ACTUATORS AND LVDT

A. General Principle

The analogy between the electrical field and the magnetic field makes the derivation of the fundamental relations in magnetic field very easy. Again the energy approach derivation will be adopted and a much more detailed discussion of this type of transducer can be found in the book by Busch-Vishniac [3].

An expression for the energy stored in a given volume of isotropic material in a magnetic field can be represented in an equation very similar to Equation (8).

$$E = \frac{\phi^2}{2p} \tag{13}$$

where ϕ is the magnetic flux and p is the permeance, or magnetic capacitance (a term similar to electric capacitance). The operation of typical inductive transducers (sensors and actuators) can be best explained by referring to the schematic diagrams shown in Figure 5. In Figure 5, two pieces of ferromagnetic rods are facing each other with the magnetic flux passing from one to the other through the air gap between them.

The air gap between the two facing electromagnetic cores shown in Figure 5 has a magnetic capacitance of almost exactly the same form as that of a parallel capacitor. If we assume the left-side element is fixed and the right-side element can move either vertically or horizontally, with the effective cross-sectional area $A = w \cdot y$, then its magnetic capacitance or permeance can be obtained as:

$$p = C = \frac{\mu A}{x} = \frac{\mu w y}{x}, \tag{14}$$

where T is the magnetic permeability of the isotropic medium.

Plug Eq. (14) into Eq. (13) and the magnetic energy stored in the air gap can be written as:

$$E = \frac{\phi 2x}{2\mu A} = \frac{\phi 2x}{2\mu wy}. \tag{15}$$

For a dynamic system, the power then can be derived by differentiating Eq. (15) with respect to time:

$$P = \frac{sE}{dt} = \frac{\partial E}{\partial \phi}\frac{d\phi}{dt} + \frac{\partial E}{\partial x}\frac{dx}{dt} + \frac{\partial E}{\partial y}\frac{dy}{dt}, \tag{16}$$

where $d\phi/dt$ is the flux rate, dx/dt and dy/dt are the velocity components of the moving magnetic element in the horizontal and vertical directions respectively. Power in the mechanical domain is equal to the product of force and velocity, and power in the magnetic domain is equal to the product of magnetic flux rate $d\phi/dt$ and *magnetomotance*, or *magnetomotive force M*. Therefore, the magnetomotive force M and force F_x in the x direction and F_y in the y direction can be derived as:

$$M = \frac{\partial E}{\partial \phi} = \frac{\phi x}{\mu A} = \frac{\phi x}{\mu wy} \tag{17}$$

and

$$F_x = \frac{\partial E}{\partial x} = \frac{\phi 2}{2\mu A} = \frac{\phi 2}{2\mu wy} = \frac{\mu wy}{2x2} M2 \tag{18}$$

and

$$F_y = \frac{\partial E}{\partial y} = -\frac{\phi 2x}{2\mu wy2} = -\frac{\mu w}{2x} M2 \tag{19}$$

The magnetomotive force in a magnetic element is a function of the magnetic field H and the length of the magnetic path, or as a function of electric current I and the number of turns in the coil according to Ampere's law:

$$M = Ni = Hl. \tag{20}$$

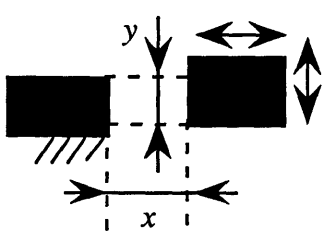


Figure 5 Basic principle of an inductive transducer.

Equation (17) represents the basic operation principle for inductive position sensing. A displacement x in the horizontal direction (with y fixed) or a displacement y in the vertical direction (with x fixed) can be transduced into a change in magnetomotance M, which can then be measured indirectly by measuring induced current i using a pick-up coil as shown in Eq. (20). Of course, a change in magnetomotance can also be achieved by changing permeability μ. This approach can also be taken in proximity sensing in similar design to the capacitive proximity sensors. Another example of using the variable permeability principle in position sensing is the widely used linear variable differential transformers (LVDTs). Detailed discussion of the LVDT and its transduction principle will be provided later in this section.

Equations (18) and (19) represent the fundamental principle for inductive actuators. They show that an electromagnetic driving force can be generated by a changing magnetomotance, that is, by supplying current to a driving coil.

In an inductive transducer (sensor/actuator), a magnetic bias must be applied. There are two ways to do this. The first way is to use a permanent magnet. The second method is to apply an electric bias, that is, to use an electromagnetic coil and supply a sinusoidal current to it. In most cases, this second approach is preferred because of its potential in securing a linear transducer relationship and the benefit to signal processing. A typical example is the linear variable differential transformer (LVDT) widely used in mechatronics applications. The advantages and wide applications of LVDT definitely justify a detailed discussion of it.

B. Operation Principle of LVDT

The schematic design of an LVDT is shown in Figure 6. One primary coil and two secondary coils are arranged as shown. A magnetic core with high permeability is inserted into the coils. At the beginning, the magnetic core is located at the middle point.

Again, we use the energy approach suggested by Busch-Vishniac [3]. Because the LVDT is symmetric, we can start the analysis by studying the left half first. When the magnetic core is in the neutral position as shown in Figure 6A, the section of the coil with air core has a magnetic capacitance:

$$C_1 = \frac{\mu_1 A}{x_0} \tag{21}$$

and the magnetic capacitance of the section with the magnetic core inserted is:

$$C_1 = \frac{\mu_2 A}{L - x_0} \tag{22}$$

where A is the cross-sectional area of the magnetic core, μ_1 is the permeability of the air core, and μ_2 is the permeability of the magnetic core. The equivalent ca-

pacitance of the combined air and magnetic cores can be calculated using the parallel rule and the total energy stored in the left-side coil can be obtained as:

$$E = \frac{1}{2C} = \phi_2 = \frac{\mu_1(L - x_0) + \mu_2 x_0}{2\mu_1\mu_2 A}\phi_2 \tag{23}$$

The magnetomotance M can be derived by differentiating energy with respect to ϕ:

$$M = \frac{[\mu_1 L + (\mu_2 - \mu_1)x_0]\phi}{\mu_1\mu_2 A}. \tag{24}$$

Now suppose the magnetic core moves from the neutral position to the right-side as shown in Figure 6B, then in the left-side, $x_0 \rightarrow x_0 + x$, and in the right-side, $x_0 \rightarrow x_0 - x$. Then the magnetomotance in the left-side of the LVDT can be obtained by substituting $(x_0 + x)$ in Eq. (24) for x_0 as:

$$M_l \frac{\phi[\mu_1 L + (\mu_2 - \mu_1)(x_0 + x)]}{\mu_1\mu_2 A} \tag{25}$$

and the magnetomotance in the right-side of the LVDT can also be obtained by substituting $(x - x)$ in Eq. (24) for x_0 as:

$$M_r = \frac{\phi[\mu_1 L + (\mu_2 - \mu_1)(x_0 - x)]}{\mu_1\mu_2 A}. \tag{26}$$

Then the magnetomotance difference of the left-side and the right-side can be obtained as:

$$\Delta M = M_l - M_r = \frac{2x(\mu_2 - \mu_1)}{\mu_1\mu_2 A}\phi. \tag{27}$$

If a sinusoidal signal is supplied in the primary coil, a sinusoidal magnetic field will be generated. This means that ϕ in Eq. (29) is also going to be a sinusoidal function of time:

$$\phi = \phi_0 \sin\omega t \tag{28}$$

where ϕ_0 is the amplitude of the magnetic flux.

Then the magnetomotance difference between the two secondary coils is also going to be a sinusoidal function of time:

$$\Delta M = M_l - M_r = \frac{2x(\mu_2 - \mu_1)}{\mu_1\mu_2 A}\phi_0 \sin\omega t \tag{29}$$

From Eq. (29), it can be seen that the magnetomotance difference in the secondary coils is of the same frequency as the current supplied to the primary coil, and its amplitude is a function of the permeability of the air and the magnetic core, the cross-sectional area of the magnetic core, the flux amplitude of the primary coil and, most importantly, the displacement x. Since all other parameters are fixed

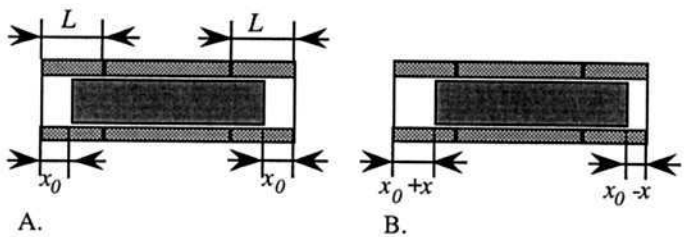


Figure 6 A schematic design of an LVDT: (A) magnetic core in neutral position; (B) magnetic core moves to the right side.

during the operation of LVDT, the amplitude of this magnetomotance difference is only modulated by the displacement to be measured. It should be noted that this position dependence is perfectly linear. The position-dependent sinusoidal signal can be demodulated easily and the position signal restored.

LVDTs have many advantages and are widely used in mechatronics products. They provide almost perfect linear responses over unlimited working range, high precision, and long mechanical life. There are many manufacturers of LVDTs and many different types of LVDTs are commercially available. There are also rotational variable differential transformers (RVDTs) available for measuring the rotational displacement.

V. ULTRASONIC SENSORS

Ultrasonic transducers are based on the measurement of sound as it travels between a transducer and an object being measured. The term ultrasonic is used to distinguish between general acoustic waves and sound in the audible spectrum (20Hz–20 kHz); most acoustic measurement is done in the ultrasonic (non-audible range) using frequencies higher than 20 kHz. Most of the ultrasonic position/displacement sensors work on the same principle as the classical sonar system. The basic design always includes an ultrasonic source (a wave generator) and a receiver, such as a microphone. There are two basic types of measurement modes: pitch-catch mode (also known as through transmission) in which a wave travels between a separate source and sensor, and pulse-echo mode in which the ultrasonic source is also used as the receiver. In the pitch-catch mode an ultrasonic wave is sent by the source and sensed by the receiver located on the target whose position is being measured. The range or distance d to the target is calculated based on the time of flight of the signal t and the speed of the ultrasonic wave propagation (c = 110,000 inches per second in air):

$$d = ct \tag{30}$$

In the pulse-echo mode the ultrasonic transducer listens for the reflected portion of the signal as it bounces off an object in its path. This echo is caused by

an acoustic impedance mismatch between the medium through which the wave is propagating, and an object located in the wave's path. Every material has a characteristic acoustic impedance z which depends upon its density and the speed at which sound propagates:

$$Z = c\Psi \tag{31}$$

Depending upon the ratio of the impedance of two materials, a portion of the sound impinging on the boundary will be reflected and a portion transmitted into the second material. More sound is reflected in cases where the impedance mismatch is larger: an impedance ratio of one means that none of the sound will be reflected back, hence no echo will be returned to the transducer. Perhaps the best known ultrasonic position sensor which operates in the pulse-echo mode is the range-finder for the auto-focusing system of the Polaroid camera [4]. It uses a single electrostatic transducer (capacitive) as both the actuator to generate the ultrasonic wave and the sensor for receiving the signal. As the exposure button is pushed, the sensor on the camera emits an ultrasonic pulse, and then listens for the return echo (caused by the impedance mismatch between the air and the object being photographed). The distance to the object is calculated as:

$$d = \frac{ct}{2} \tag{32}$$

and the camera is then properly focused before exposing the film. The range-finder in the Polaroid camera uses a group of wave pulses at different frequencies to make sure the distance measurement is accurate.

In addition to performing range-finding measurement in air, ultrasonic sensors can be sued to evaluate position in solid materials. As in the air, both pitch-catch and pulse-echo modes can be used to measure positions. Pulse-echo, however, is more popular for the measurement of position in solids. In many testing applications, such as the detection of cracks in a reactor wall, there is access to only one side of the solid material. In such a case, the ultrasonic wave will bounce off any cracks or voids in the material, as well as the far boundary of the solid. Each reflector will be detected separately by the transducer.

The accuracy with which position can be measured ultrasonically depends upon the frequency of the signal. In general, measurement resolution is determined by the signal wavelength Σ which is in turn determined by the signal frequency:

$$c = f\Sigma \tag{33}$$

where f is the frequency.

This would imply that high frequencies should be used for range-finding in order to maximize the accuracy of the measurement. Unfortunately frequency cannot be increased without bound, due to acoustic attenuation which increases as a function of frequency. For instance, a 1 MHz signal will only travel a few inches in air before it becomes indistinguishable from the background noise. A good treatment of issues pertaining to ultrasonic measurement is provided by Krautkramer and Krautkramer [8].

VI. FIBER-OPTIC SENSORS AND OPTICAL INCODERS

A. Fiber-Optic Sensors

Fiber-optic sensors have some significant advantages: for example, freedom from electromagnetic interference, flexibility, low signal attenuation, rugged structure and smaller physical sizes compared with conventional optical system, and potentially high accuracy. Probably the most important fiber-optic position sensors is the fiber-optic interferometer. Compared with a conventional interferometer, a fiber-optic interferometer has the advantages of, for example, much smaller physical size, stable performance (less sensitive to environmental conditions such as vibrations), and ease of assembly.

There are two main categories of fiber-optic sensors. The first category is intrinsic, and the second category is extrinsic. In an intrinsic fiber-optic sensor the transmission properties of the optic-fiber are modulated by the physical parameter to be measured. For example, if a piece of optic-fiber coated with magnetostrictive material can be used for magnetic field detection because a change in magnetic field can cause an extension in the optic-fiber, this extension of fiber can then cause a change in the total length of the light-path and can be measured by measuring the phase difference of the light signal using the interferometric principle. Similarly, a fiber-optic temperature sensor can also be designed by introducing another transduction mechanism to transform the temperature change into a change in the length of an optic-fiber.

In extrinsic types of fiber-optic sensors, the optic-fiber is only used to transmit the light signal. A typical example is a proximity sensor that can be built with a photoelectric sensing cell and an LED pig-tailed with optic-fibers. In this case, the optic fibers are used as the sensing head. Compared with a regular proximity sensor with photoelectric sensor and LED, this type of sensor has unique advantages in applications that require position sensing in limited space.

B. Optical Encoders

Optical encoders are high-precision rotational position sensing devices that are widely used in mechatronics products, especially in industrial robots. They are mounted on a rotary shaft to generate a digital signal to sense its rotational position. There are two types of optical encoders: absolute and incremental, see Kuo [9]. In an absolute encoder, multiple concentric rings of binary code are etched on a code-wheel. The code along any radial line on this code-wheel represents an absolute value of the rotational position. One light source and one light detector are used to read the code. Using absolute encoders, it is necessary to know only one radial line of code to determine the position of the shaft. Absolute encoders can be very accurate. Angular position can be measured to an accuracy of one part in 2^{10}. The incremental optical encoders are simpler and cheaper compared with

the absolute ones. Two pairs of photoelectric sensor and light sources are used to sense both the direction and size of the motion. There are two optional designs. One uses a one-track code-disk and another uses a two-track (one inner and one outer track) code-disk. When a two-track disk is used, each track on the code-disk requires one pair of photoelectric sensor and light source. When a one-track disk is used, the two photoelectric sensors are arranged in such a way that one is displaced from the other by one and one-half slot widths. The pulse trains from the photoelectric sensors are then counted and interpreted by an electronic circuit to obtain the rotational position and direction.

VII. PHOTOELECTRIC SENSORS AND LATERAL-EFFECT PSD

A. Photoelectric Sensors

Photoelectric sensors are another major category of sensor that is used in proximity and position sensing. The commonly used photosensors include photoelectric tubes, photodiodes, phototransistors, photoconductive transducers, photovoltaic cells, CCDs, and lateral-effect position sensitive detectors (PSDs). In this section, we will briefly introduce the general principles of position and proximity sensing with photoelectric sensing cells, and then spend most of the time discussing the lateral-effect position detectors (PSDs) because of their advantages in providing highly precise, highly linear position measurements.

Most of the photoelectric sensors can only provide an output electric signal dependent on the total power of light falling on the sensing cells. In general, there are two ways to use them for position sensing as shown in Figure 7. The first method is to arrange the light source (for example, light emitting diodes, LEDs), and the photoelectric sensor in opposite positions as shown in Method A of Figure 7. By moving into or out of the light beam, the position, presence, or absence of an object can be sensed. Another common method of using photoelectric cells in position sensing is shown in Method B of Figure 7. In this arrangement, the light source and the photo-sensor are located in the same plane and normally fabricated in one package. Many manufacturers supply this type of proximity sensor. They are supplied by many manufacturers at very low prices. The moving object whose position is being sensed reflects a light beam back onto the surface of the photoelectric sensing cell. Typically, this type of proximity sensor has a larger working range and lower precision compared with the capacitive sensors. The precision may be increased by reducing the measurement range.

Silicon image sensors such as charge-couple devices (CCDs) belong to another important category of photoelectric sensors that have very wide mechatronics applications. The CCD can be used to transfer an optical image or frame from light-sensitive array to a digital output. It has become popular in the manufacturing industry because of the wide applications of microcomputers. The major disadvantage of the CCD cameras is that they are quite expensive.

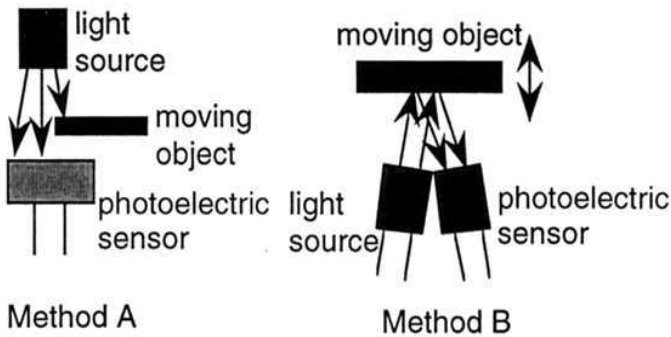


Figure 7 Two methods to use the photoelectric cells for position measurement.

B. Lateral Effect Position Sensitive Detector (PSD)

The lateral-effect position sensitive detector (PSD) is widely used for high precision position measurement. This type of position detector has some prominent advantages compared with other optical and photoelectric sensors. It can measure displacement in a spatially continuous manner, unlike other types of large sensitive area detectors such as charge-coupled devices (CCDs). It provides either one-dimensional or two-dimensional, highly linear, spatially continuous, fast, real time position/displacement measurement. Submicron resolution over a large working range can be obtained. These advantages make this type of position sensor very suitable for mechatronics applications that require very high precision position/displacement measurement over a large working space.

Operation Principle and Different Types of PSDs

It is commonly known that a light beam projected onto the surface of a p-n junction produces a photopotential on each plane of the junction. The photopotential will induce photocurrent if there are electrodes on the boundary of a junction plane. The induced photocurrent will flow laterally toward the electrodes on the boundary because of the photopotential gradient in the lateral direction. This phenomenon is called the lateral photoeffect and was first discovered by Schottky [10] in 1930. The discovery did not draw enough attention until 'rediscovered' by Wallmark [11] in 1959. Wallmark designed a device composed of a Ge-In p-n junction in which a light beam projected onto the surface causes a position dependent photopotential difference between the point contacts on the sensor surface.

Locovsky [12] derived the fundamental equations describing the diffusion and recombination processes for the steady-state and the small-signal transient cases for a modern PSD. Connors [13] investigated the one-dimensional model of a reverse-biased PSD. Woltring [14] extended Connors' analysis to the two-dimensional rectangular PSD operated in both the unbiased (small signal) and the

fully reverse-biased modes. The fundamental equation for the potential distribution in a PSD that is in steady-state and fully reverse-biased provided by Woltring [14] has a very simple form as follows:

$$\frac{\partial^2 U(x,y)}{\partial x^2} + \frac{\partial^2 U(x,y)}{\partial^2 y} = -\frac{\rho}{w} I_s(x,y) \tag{34}$$

where $U(x,y)$ is the electric potential at point (x,y) generated by the light beam projected on the sensor surface, Ψ the surface resistivity, w the thickness of the resistive layer of the sensor surface, and $I(x,y)$ the generated photocurrent density. Using this equation and boundary conditions set by the design of electrodes, the potential distribution can be solved either analytically or numerically. Then the photocurrents flowing to electrodes can also be obtained. Woltring [14] provided a very detailed analysis for two types of the most commonly used PSDs: the duolateral and tetralateral, whose geometries are shown in Figures 8A and 8B. Wang and Busch-Vishniac [15] studied the linearity for all three types of PSDs (duolateral, tetralateral, and pin-cushion) commercially available today, and proposed an alternative, clover geometry design.

In the schematic diagrams shown in Figure 8, the shaded area is the effective (useful) device area, and the dark areas the electrodes. A duolateral, duoaxis PSD has two electrodes on each side of the p-n junction. A tetralateral type PSD has all four electrodes on one side of the p-n junction. A pin-cushion type PSD shown in Figure 8c has curved boundaries and the effective transducer area is formed by a rectangle whose sides are tangent to the innermost points on the edges. The electrodes are point contacts located at the four corners.

To have the best performance and prevent any possible recombination across the p-n junction upon illumination, it is always preferable to operate PSDs in the fully reverse-biased mode. In the fully reverse-biased mode, the photocurrent generated by a falling light beam distributes to electrodes on the boundaries of the sensor surface. These photocurrents are then measured for position sensing of the lightbeam. Full reverse-biasing of lateral-effect PSDs can increase the thickness of the depletion layer of the p-n junction, and reduce the current flowing across the p-n junction to almost zero, therefore preventing signal loss due to surface recombination. It also reduces the effective capacitance of the p-n junction, therefore increasing the response speed of the PSD.

In the fully reverse-biased mode, the duoaxis duolateral PSD is inherently linear because the x and y contacts are on opposite sides of the junction. Compared to the duolateral PSD, the tetralateral PSD with four extended ohmic contacts (electrodes) on one side of the p-n junction has disadvantages, namely, the electrode structure makes cross-talk and nonlinearity inevitable. In a tetralateral type PSD, the generated photocurrent is divided into four parts instead of two parts as in a duolateral type PSD, so that the resolution is about half that of the duolateral one for the same noise level. However, the tetralateral PSD also has advantages when compared to the duolateral PSD: a faster response, a much lower dark current, an easier reverse-bias application, and a lower fabrication cost. The pin-

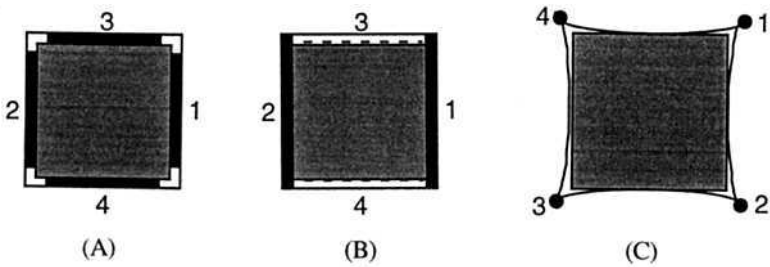


Figure 8 (A) Tetralateral position sensitive device. All four circuits are on a single surface. Shaded area is photosensitive, and the heavy solid lines indicate electrodes. The effecctive device area is shown hatched; (B) duolateral duoaxis position sensitive device. Two of the electrodes are on the p-surface and two on the n-surface; (C) Pin-cushion type position sensitive device. Curved boundary with four electrodes located in corners.

cushion type of PSD by Hamamatsu [16] has a performance somewhere between those of the duolateral and tetralateral ones.

All three types of PSDs can provide the cartesian two-dimensional location of a light spot. Geometries suitable for other coordinate systems have been studied, including work by Xing and Boeder [17] on a PSD for measurement of angular position.

Signal Processing for PSD

When a light beam scans across the surface of a two-dimensional PSD, the photocurrents flowing to the four electrodes will change as a function of the light spot position. A power fluctuation of the light beam or variations of environmental illumination may also cause a change in the total photocurrent generated, therefore resulting in changes of photocurrents flowing to electrodes. A normalization technique is always adopted to eliminate the influences of beam intensity or environmental illumination on the position signals. Suppose I_1, I_2, I_3, and I_4 are the photocurrents from PSD electrodes as numbered in Figure 8; then the normalized x and y position signals for a duolateral or a tetralateral PSD are, see Woltring [14]:

$$x = K_x \frac{I_1 - I_2}{I_1 + I_2}, \tag{35}$$

$$y = K_y \frac{I_3 - I_4}{I_3 + I_4}, \tag{36}$$

where K_x and K_y are the amplification factors of the x and y channels in the signal processing circuit.

Similarly, the normalized x and y position signals for a pin-cushion type PSD are, see Hamamatsu [16]:

$$x = K_x \frac{(I_1 - I_2) - (I_3 - I_4)}{I_1 + I_2 + I_3 + I_4}. \tag{37}$$

$$y = K_y \frac{(I_1 + I_4) - (I_2 + I_3)}{I_1 + I_2 + I_3 + I_4}. \quad (38)$$

Generally, the photocurrents from the electrodes of a PSD are amplified with preamplifiers first, then additions, subtractions and divisions are carried out in the following stages of the signal processing circuit.

The resolution of a position sensing system using a PSD as a sensing cell is ultimately limited by the signal to noise ratio (S/N) of the system. The signal can be boosted by increasing the power of the light beam. However, the intensity of the light beam is limited by the saturation level. The other option is to reduce the total noise of the system. To reduce the noise level would not be an easy task because it is very hard to separate the signal from the noise without using modulation techniques. This means that, if further improvement of the signal to noise ratio is desired, a modulation technique has to be used.

Different Ways to Use PSDs

Because PSDs can only detect the relative movement of a light spot on its surface, to use it for noncontact position sensing it is necessary to convert movement of the object into the relative movement of a light spot on the sensor surface. There are three ways to accomplish this as shown in Figure 9. The first method is to fix the PSD on the moving object to be sensed while keeping the light source stationary. The second method is to fix a light source such as an LED or a semiconductor laser on the moving object to be measured while holding the PSD stationary. Both approaches require tethering power wires to the moving object, either supplying to PSDs or to the light sources. In some applications, these wires may not be accepted. The third method is to fix a rigid, opaque plate with a pin-hole onto the object to be sensed while both a highly collimated light source and a PSD are kept stationary. The light emitted from the light source is blocked by the opaque plate except at the pin-hole. When the object being sensed moves around, a different part of the light beam passes through the pin-hole, causing a light spot to move on the PSD surface.

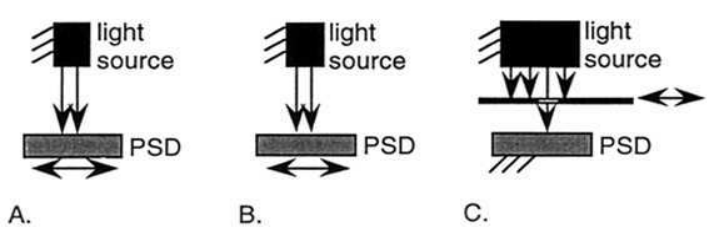


Figure 9 Three ways to use PSDs for noncontact position sensing: (A) light source fixed; (B) PSD fixed; and (C) both PSD and light source fixed.

One major advantage with the third method is that there is no cross-talk between the motion within the $x - y$ plane and the out-of-plane rotation. This is possible because a PSD detects the location of the centroid of the light spot falling on its surface. When the object has an out-of-plane rotation, the projection of the pin-hole on the PSD surface becomes an ellipse instead of a circle, but both have the same centroid if the light beam is of uniform intensity. Consequently only the total amount of light falling on the surface of the PSD is decreased. This does not affect the output position signals because they are normalized by the signal processing circuit. The disadvantage of the third approach is that the measurement accuracy depends heavily on the uniformity of the light source. Ideally, we would like to use a highly collimated light source, like a laser.

Modulation Technique for PSD

The essence of modulation technology is to narrow the frequency bandwidth of the transmitted signal, so that all other background noises can be eliminated with appropriate filtering. This approach, therefore, can dramatically improve the signal to noise ratio of a sensing system. There are many kinds of modulation methods. The two most popular methods used in radio communication are frequency modulation (FM) and amplitude modulation (AM). Both the frequency modulation and amplitude modulation of a continuous light beam are not suitable because it would be very difficult to have a light beam whose intensity is a sinusoidal function. Other alternative options include pulse modulation approaches. In pulse modulation, the carrier signal is a pulse train instead of a sinusoidal wave. There are three kinds of pulse modulation approaches commonly in use. One is called pulse position modulation (PPM), where the pulse position is modulated by the signal to be transmitted. This method is not suitable because it is very difficult to design a measurement scheme in which the pulse position can be modulated by the displacement of a measurement object. The second method is called pulse duration modulation (PDM) where the pulse duration must be modulated by the moving object, which is very difficult to implement. The third method is the pulse amplitude modulation (PAM), that is, the pulse amplitude is modulated by the signal to be transmitted. This method is recommended by the manufacturer of the PSD, Woltring [14], and has been used by Wang and Busch-Vishniac [18], and a resolution down to 0.1 micrometers has been achieved over a large work space. Interested readers are referred to the reference list at the end of this chapter.

The principle of PAM is explained in Figure 10. Shown in Figure 10A is a pulse train signal supplied to a light source, such as an LED, to generate a light beam which switches on and off continuously at the modulation frequency. If a light beam such as an LED is fixed on a moving object whose motion is being monitored while the PSD is stationary, then the photocurrents flowing to the PSD electrodes are also pulse trains at the same frequency as the input light beam. Therefore the amplitudes of the photocurrents arriving at the electrodes are modulated functions of the object movement. Two synchronized delayed pulse trains with the same frequency shown in Figure 10B and Figure 10C are used to

trigger the sample/hold amplifiers to restore the DC mode. One has its rising edge corresponding to the light beam on period. The other has its rising edge located at the light beam 'off' period. The output signals are then subtracted from each other to obtain the amplitude of the position signal. Detailed design of the circuit is not able to be provided here. Interested readers may find more information about the circuit design in the paper by Wang and Busch-Vishniac [18].

Measuring Multiple Beams with One PSD

Normally, one PSD is needed to measure the position/displacement of one light beam. However, if an optical measurement system uses multiple beams and multiple detectors, the system hardware can become large and redundant. Because of the cost of the PSDs, the system cost may become prohibitively high. Further, alignment difficulties increase dramatically as more PSDs are used. Finally, if several PSDs are used in a sensing system and each one of them needs to be calibrated separately, calibration of the sensing system will demand significant efforts. Therefore it is highly desirable to use a single PSD to measure the positions/displacements of multiple light beams instead of one SD for each light source. The resulting advantages include: a more compact system, lower cost, faster calibration, and preservation of almost identical physical and environmental conditions for every sensing element.

A suitable modulation technique which permits the measurement of multiple light spots using a single optical sensor has been provided by Tian et al. [19]. Borrowing from technology widely used in the telecommunication industry, a modulation method is used to permit one sensor to monitor multiple light sources. Each light source is modulated at a different frequency. When there are multiple beams irradiating a single PSD the total photo-voltage generated will equal the sum of the photo-voltages generated by all beams if the PSD is working in the linear region, i.e., if the reverse-biased voltage is large enough to prevent recombination. The

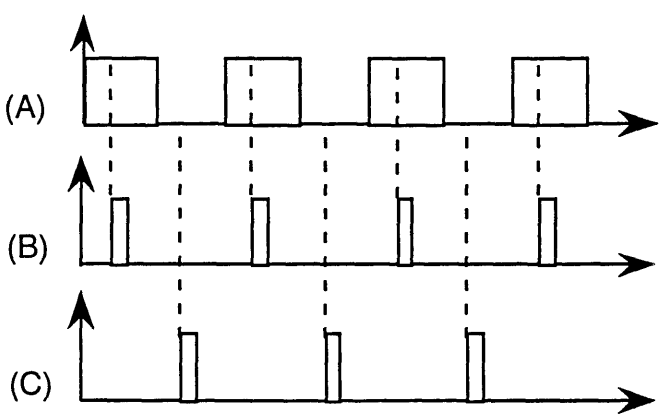


Figure 10 Pulse amplitude modulation technique for position sensing with PSD.

position-sensitive photocurrents from the PSD sensor are then superpositions of the photocurrents generated by all light beams. The position/displacement information on a specific light beam can be obtained through demodulation. An alternative mean of accomplishing separation of signals from multiple light sources would be to switch light sources on and off sequentially. While this method may be acceptable for some applications, it is generally less desirable than a frequency separation approach because the time separation method introduces transients into the system. In order to prevent the system transients from causing saturation, the signal level must be lowered, resulting in poorer sensor measurement accuracy. Tian et al. [19] reported the successful implementation of the idea and very high resolutions for displacement measurement of multiple light beams.

The number of light beams the PSD can measure is only limited by two factors: signal loss from recombination and the bandwidth of the PSD.

REFERENCES

1. S Middelhoek, AC Hoogerwerf. Classifying solid-state sensors: the sensor effect cube. Sensors and Actuators 10:1,1986.
2. DC Karnopp, DL Margolis, RC Rosenberg. *System Dynamics.* 2nd. ed. New York: Wiley & Sons, 1990.
3. IJ Busch-Vishniac. *Electromechanical Sensors and Actuators.* New York: Springer-Verlag, in press.
4. Polaroid Ultrasonic Ranging Experimenter's Kit. Cambridge, Mass: Polaroid Corp., 1980.
5. RC Luo. Sensor technologies and microsensor issue for mechatronics systems. IEEE/ASME Transactions on Mechatronics 1(1):39–49, 1996.
6. YS Lee, KD Wise. A batch-fabricated silicon capacitive pressure transducer with low temperature sensitivity. IEEE Transactions on Electron Devices 29(1):42–48, 1982.
7. D Bosch, B Heimhofer, G Muck, H Seidel, U Thumser, W Welser. A silicon microvalve with combined electromagnetic/electrostatic actuation. Sensors and Actuators A 37–38: 684–692, 1993.
8. J Krautkramer, H Krautkramer. *Ultrasonic Testing of Materials.* 3rd ed. Berlin: Springer-Verlag, 1983.
9. BJ Kuo. *Automatic Control Systems, 4th ed.* New York: Prentice-Hall, 1982.
10. W Schottky. Ueber den Entsstehungsort der Photoelektronen in Kupfer-Kupferoxydul-photozellen. Phys. Z. 31:913–925, 1930.
11. T. Wallmark. A new semiconductor photocell using lateral photoeffect. Proc. IRE 45:474–483, 1957.
12. G Lucovsky. Photo-effects in nonuniformly irradiated p-n junctions. Journal of Applied Physics 33:1088–1095, 1960.
13. P Connors. Lateral photodetector operating in the fully reverse-biased mode. IEEE Transaction on Electron Devices 18:591–596, 1971.
14. HJ Woltring. Single- and dual-axis lateral photodetectors of rectangular shape. IEEE Transactions on Electron Devices 22:581–586, 1975.
15. W Wang, IJ Busch-Vishniac. The linearity and sensitivity of lateral effect position sensitive devices—an improved geometry. IEEE Transactions on Electron Devices 36:2475–2480, 1989.
16. Hamamatsu Photonics Corp. Position Sensitive Detectors. Product Catalogue, 1987.
17. YZ Xing, CPW Boeder. A new angular-position detector utilizing the latreal photoeffect in Si. Sensors and Actuators 7:153–166, 1985.
18. W Wang, IJ Busch-Vishniac. A four-dimensional non-contact sensing system for micro-automation machines. The Sensors and Instrumentation for In-Process Monitoring of Manufacturing Technical Session, ASME Winter Annual Meeting, Dallas, Texas, November 25–30, 1990.

19. D Qian, W Wang, IJ Busch-Vishniac, AB Buckman. A method for measurement of multiple light spot positions on one position-sensitive detector (PSD). IEEE Transactions on Instrumentation and Measurements 42(1):14–18, 1993. The preliminary results were also presented at the IMTC'92, New York City, New York, May 1992.